MIE 1624 Introduction to Data Science and Analytics
Assignment 3

Name: Ziruo Song
Wednesday, December 7, 2022

Background and Objective

For a new program named “Master of Business and Management in Data Science and Artificial Intelligence” at University of Toronto with focus on both technical and business skills, a course curriculum with a logical sequence of 8 courses including 3 to 8 topics for each are developed based on clusters of skills that are necessary for data scientist or analyst.

Part 1. Data collection and basic cleaning

The unique 1427 job postings are scarped from Indeed, with the setting of the geographical locations to be Canada or the US, and the job titles to be data scientist or analyst. Apart from the basic informations, like company, rating, salary, the full descriptions are cleaned with the removal of punctuations and stop words, and lemmatizations are performed on the cleaned full descriptions, and stored to the file named “*webscraping_results_assignmnet3.csv” with the basic informations scraped.* As the noise during the exploration of the intended meaning of the text can be reduced by removing the inflectional endings and returning the base form of a word, lemmatization is a technique widely used in Natural Language Processing.

Part 2. Exploratory data analysis and feature engineering

Part 2.1 Extract skills and further cleaning

Ten soft/business including communication, presentation, teamwork, and fourteen hard/technical skills, like excel, python, r, java, or tableau are defined, and each of them is extracted from the full description after lemmatizations and consider as a feature labeled with 1 if being mentioned in the full job description, and 0 otherwise. Only title, company, location, salary and lemma along with each skills are kept for the further explorations, and the data type is transformed to string or numeric according to the each feature.

Part 2.2. Visualization 1-Word Cloud

In the next step, three word clouds of the full job descriptions after lemmatization are performed, based on both or either of the two titles, but the less informative words who appeared much often, like ‘data’, ‘experience’, ‘work’, etc, are removed in this part to reduce noises. For the rest, we could see ‘business’, ‘team’ , ‘management’ , ‘analysis’ were mentioned quite often regardless of the job titles, shown in ***Figure 2-1***. Apart from common words for both job titles, for job postings about data scientist, ‘statistical’, ‘technology’, ‘python’, ‘machine’, ‘insight’, ‘process’, and ‘modelling’ are much frequent according to ***Figure 2-2***. On the contrary, ‘report’, ‘information’, ‘service’ , ‘sql’, ‘product’ can be caught more frequent in the descriptions of analyst by *Figure 2-3*. The visualization by word clouds can help us to grasp the frequency or importance of the words, but be careful with the noise.

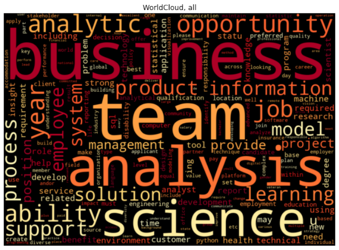


Figure 2-1, Data Scientist & Analyst

Figure 2-2, Analyst

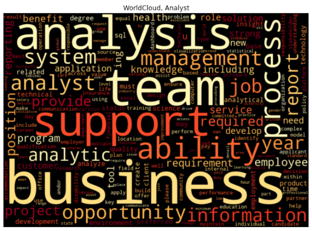


Figure 2-3, DS

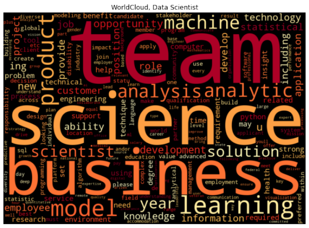


Figure 2-4, Hard skills

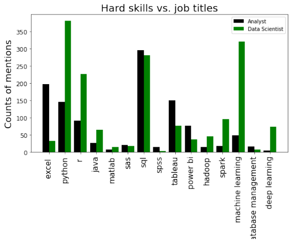
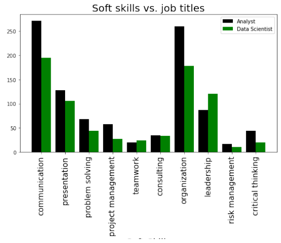


Figure 2-5, Soft skills



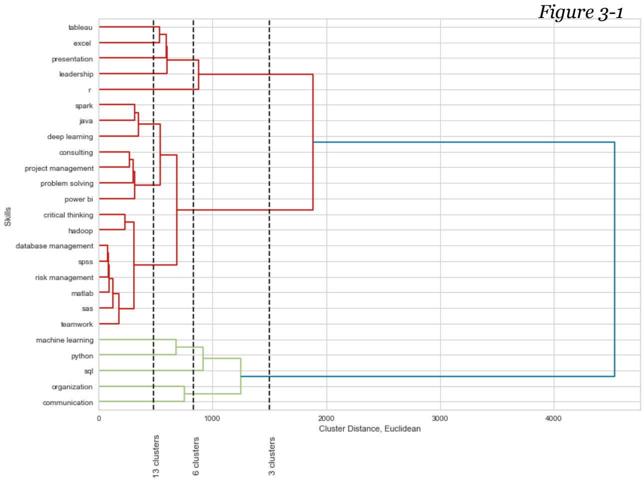
Part 2.2. Visualization 2- Bar plots.

Additionally, the importances of hard and soft skills for data scientist and analyst are compared in the bar plots. By the comparison between different hard skills in ***Figure 2-4***. The use of Python, R, or machine learning seems to be more necessary to do a decent job as data scientist, but the data visualization tools, like Excel, Tableau, Power BI should be equipped by an analyst more than data scientist, as the counts of mentions of these hard skills are higher. However, the requirement for SQL may be valued similarly for both of the two jobs. By the top-right plot of ***Figure 2-5***, most of the soft skills in our list are required slightly more for analyst compared with data scientist, except leadership and team work, but the level of importance are quite similar for the these two jobs. The comparisons in bar plots can helps us to understand the difference in the importance of each skill for different job titles, but note the extractions of the skills may not be accurate sometimes, which can be attributed to the difference in form of the words.

Part 3. Hierarchical clustering implementation.

Next, the Ward hierarchical clustering with Euclidean distance is performed, and a dendrogram resulted from the algorithm is displayed below. 6 clusters are selected based on both the common understandings of these topics, and the importance of each of them by Part 2.

Figure 3-1

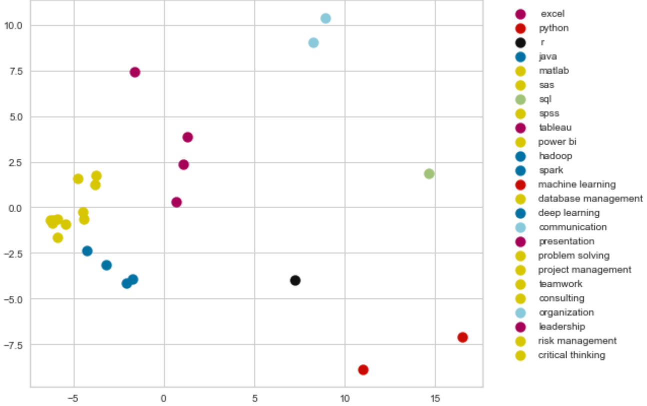


As we want to design 8-12 course with 3-8 topic included in each course, we could select the several topics from the same cluster and groups them into a course. As the skills used in each course are not limited to be unique, and the important ones (observed from the bar plots of Part 2) can be the focuses of several courses, which can enable the students to grasp the sufficient knowledge on the topic and learn how to apply the skills well. For example, MIE1624 and many other courses consider Python as the major language. On the contrary, for the topics that appear not so frequent, like SAS, Matlab or SPSS, which are grouped to a single large cluster, representing it is hard for the algorithm to separate them which need more considerations for the design of the course curriculum. One try could be aggregating some of them into some introductory courses which provide the student with the general but basic understanding of these topics. Note this algorithm involves lots of arbitrary decisions so works not well on very large data sets.

Part 4. K-means clustering implementation.

With this algorithm, we consider skills as indexes and each sample as a feature, then principle component analysis is performed to reduce the dimensions to 2, enabling the k-means to work, and KElbowVisualizer is used to determine the optimal number of clusters to be 7. The default metric here is the sum of squared distances from each point to its assigned centre. Note K means is dependent on initial values using elbow method. The results indicated on the ***Figure 4-1*** is similar with ***Figure 3-1*** using Hierarchical clustering, but there is no obvious big cluster.

Figure 4-1



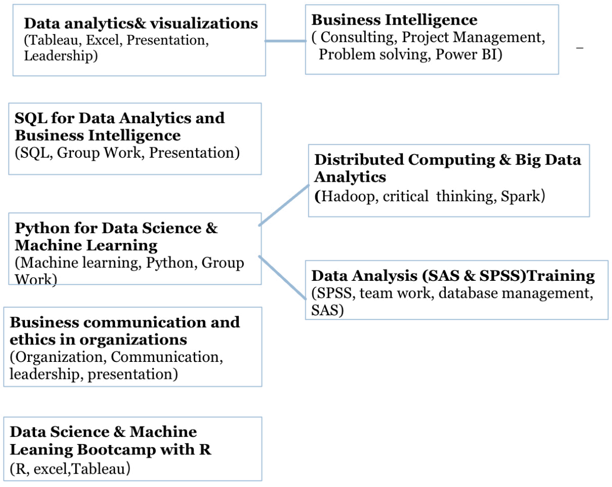
Part 5. Interpretation of results, discussion and final course curriculum

Apart from the observations by the two algorithms mentioned above, consider the most important skills derived from part 2 with priority, and take the pre-condition before learning some skills into account at the same time.

Prerequisites of these skills:

- Spark: Prefer intermediate programming experience in Python or Scala.
- Power BI: Understand Excel, basic data analytics and business analytics concepts
- SPSS: Python, which makes it much easier to do complicated programming.

Prerequisite course are indicated by the left ones being connected.



Data analytics& visualizations
(Tableau, Excel, Presentation, Leadership)

Business Intelligence
(Consulting, Project Management, Problem solving, Power BI)

SQL for Data Analytics and Business Intelligence
(SQL, Group Work, Presentation)

Distributed Computing & Big Data Analytics
(Hadoop, critical thinking, Spark)

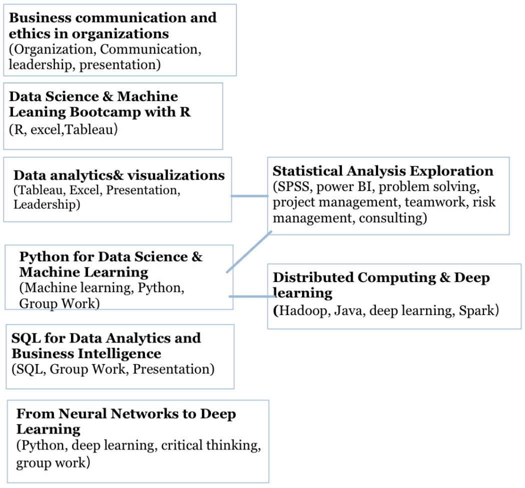
Python for Data Science & Machine Learning
(Machine learning, Python, Group Work)

Data Analysis (SAS & SPSS)Training
(SPSS, team work, database management, SAS)

Business communication and ethics in organizations
(Organization, Communication, leadership, presentation)

Data Science & Machine Leaning Bootcamp with R
(R, excel,Tableau)

Figure 5-1, by Hierarchical clustering



Business communication and ethics in organizations
(Organization, Communication, leadership, presentation)

Data Science & Machine Leaning Bootcamp with R
(R, excel,Tableau)

Data analytics& visualizations
(Tableau, Excel, Presentation, Leadership)

Statistical Analysis Exploration
(SPSS, power BI, problem solving, project management, teamwork, risk management, consulting)

Python for Data Science & Machine Learning
(Machine learning, Python, Group Work)

Distributed Computing & Deep learning
(Hadoop, Java, deep learning, Spark)

SQL for Data Analytics and Business Intelligence
(SQL, Group Work, Presentation)

From Neural Networks to Deep Learning
(Python, deep learning, critical thinking, group work)

Figure 5-2, by K means Clustering

There is no obvious difference, but the result by K-means would be selected, as the big cluster after the implementation of Hierarchical clustering is split to smaller sub-clusters here, which means the algorithm can learn those skills slightly better. There are also less courses requiring pre-requisition which may provide more flexibility in course selection. The results may be enhanced by getting more skills involved.